DisDP: Robust Imitation Learning via Disentangled Diffusion Policies

Anonymous authors
Paper under double-blind review

Keywords: Imitation learning, Diffusion policy, Multi-View Disentanglement

Summary

This work introduces Disentangled Diffusion Policy (DisDP), an Imitation Learning (IL) method that enhances robustness by integrating multi-view disentanglement into diffusion-based policies. For robots to be deployed on a large scale across various applications they have to be robust against different perturbations, including sensor noise, complete sensor dropout and environmental variations. Existing IL methods struggle to generalize under such conditions, as they typically assume consistent, noise-free inputs. To address this limitation, DisDP structures sensory inputs into shared and private representations, preserving task-relevant global features while retaining distinct details from individual sensors. Additionally, Disentangled Behavior Cloning (DisBC) is introduced, a disentangled Behavior Cloning (BC) policy, to demonstrate the general applicance of disentanglement for IL. This structured representation improves resilience against sensor dropouts and perturbations. Evaluations on The Colosseum and Libero benchmarks demonstrate that disentangled policies achieve better performance in general and exhibit greater robustness to any perturbations compared to their baseline policies.

Contribution(s)

1. *Introducing Disentangled Diffusion Policy (DisDP), an Imitation Learning (IL) method that improves robustness to sensor noise and dropouts by structuring sensor inputs into shared and private representations.*

   Context: Prior IL methods rely on consistent, noise-free sensor inputs, which limits their effectiveness in real-world scenarios.

2. *Leveraging Multi-View Disentanglement for enhancing robustness and interpretability of the behavior policies.*

   Context: This implementation displays the general application of Multi-view disentanglement for robot policies. It uses contrastive and orthogonality constraints to separate shared and unique information. This approach enhances the interpretability by visualizing the shared and private representations.

3. *Provides an extensive experimental analysis on the effect of the Colosseum and Libero benchmarks in sensor failure and environment perturbation scenarios.*

   Context: This work shows the performance degradation of behavior policies with unreliable sensors and environmental variations. Additionally, it shows how disentangled latent spaces benefit in these scenarios.

DisDP: Robust Imitation Learning via Disentangled Diffusion Policies

Anonymous authors
Paper under double-blind review

Abstract

This work introduces Disentangled Diffusion Policy (DisDP), an Imitation Learning (IL) method that enhances robustness. Robot policies have to be robust against different perturbations, including sensor noise, complete sensor dropout and environmental variations. Existing IL methods struggle to generalize under such conditions, as they typically assume consistent, noise-free inputs. To address this limitation, DisDP structures sensors into shared and private representations, preserving global features while retaining details from individual sensors. Additionally, Disentangled Behavior Cloning (DisBC) is introduced, a disentangled BC policy, to demonstrate the general applicance of disentanglement for IL. This structured representation improves resilience against sensor dropouts and perturbations. Evaluations on The Colosseum and Libero benchmarks demonstrate that disentangled policies achieve better performance in general and exhibit greater robustness to perturbations compared to their baseline policies.

1 Introduction

For robots to be deployed on a large scale across various applications, they have to be robust against different perturbations, including environmental variations, sensor noise, and complete sensor modality dropout. Sensor modality dropout refers to the unavailability of sensors during inference, that have been available during training. While current research has explored environmental variations, and perturbations in behavior learning scenarios Pumacay et al. (2024), sensor modality dropout remains an understudied challenge. Current methods that try to solve sensor modality dropout struggle on challenging datasets Skand et al. (2024); Hao et al. (2023b). To tackle this challenge, we propose **Disentangled Diffusion Policy (DisDP)**, a method that disentangles the latent space of different sensor modalities into shared and private embeddings.

Integrating information from multiple sensors has shown to improve the robust and adaptive performance of the learned policies, especially in scenarios where a single sensor may be insufficient or unreliable (Li et al., 2022; Shridhar et al., 2023; Thankaraj & Pinto, 2023; Liu et al., 2024c; Jones et al., 2025). However, multi-sensor setups are prone to a variety of challenges. Sensors can introduce noise due to calibration errors, hardware or environmental conditions. Additionally, sensor failures or data dropouts can occur due to hardware malfunctions. Most of the current robot learning approaches depend on consistent noise-free sensory input during inference(Shridhar et al., 2023; Thankaraj & Pinto, 2023; Reuss et al., 2024b). This dependency limits their effectiveness in handling noisy sensors and sensor dropouts as shown in Section 4. This work focuses on image modalities and the influence on robot policies, when one or more sensors are unreliable or unavailable, as illustrated in Figure 1. Among the various robot learning approaches, Imitation Learning (Argall et al., 2009; Osa et al., 2018) has emerged as a widely used method for teaching robots complex behaviors through expert demonstrations. SOTA multi-task IL methods demonstrated strong performance on diverse robot learning tasks (Pari et al., 2021; Shafiullah et al., 2022; Shridhar et al., 2023; Hao et al., 2023b; Reuss et al., 2024b; 2023; Chi et al., 2023; Donat et al., 2025). Despite

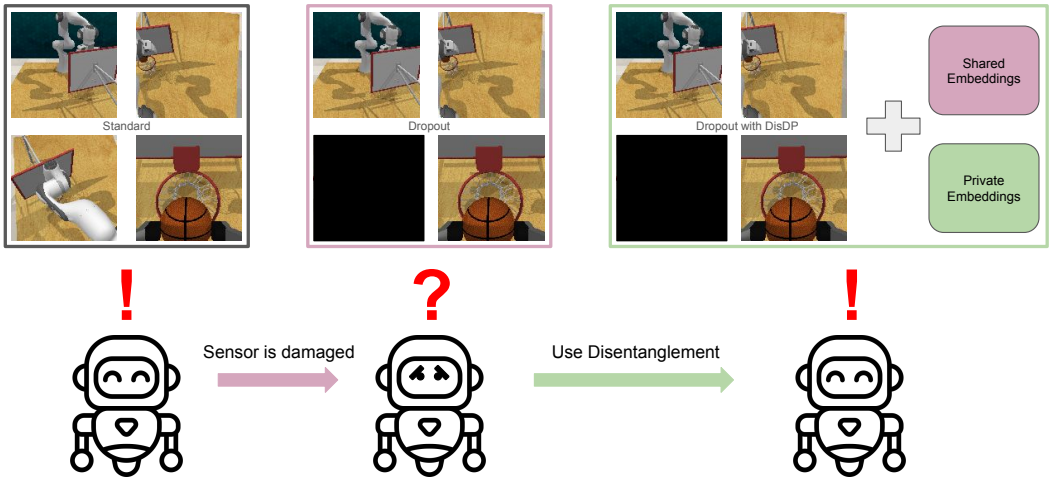

Figure 1: Robotic policies depend on multiple sensory inputs, making them susceptible to sensor failures. This work investigates how disentangling sensory information into shared and private embeddings can enable robust policy learning under sensor dropouts.

these advancements, most IL methods rely on latent representations not explicitly designed to handle noisy or missing sensory data, making them vulnerable to sensor degradation or dropout.

Therefore, the contribution of this paper is the introduction of disentanglement for IL policies to enhance robustness. Sensor inputs are separated into shared and private embeddings, enhancing robustness and interpretability of the policy. The key issues addressed include added noise in camera inputs, camera modality dropouts and environmental perturbations. The disentanglement is applied on a score-based diffusion policy for robot action prediction (Reuss et al., 2023; 2024b), as well as traditional BC using a Transformer backbone (Mandlekar et al., 2021a). The extensive experimental analysis shows, that disentangling the latent space improves overall performance and reduces performance loss under unreliable sensors and environmental perturbations.

2 Related Work

2.1 Multi-modal Imitation Learning

Imitation Learning (IL) (Argall et al., 2009; Osa et al., 2018) has demonstrated strong performance across diverse robotic manipulation tasks (Zitkovich et al., 2023; Octo Model Team et al., 2024). In particular, diffusion models (Ho et al., 2020; Karras et al., 2022) have emerged as effective IL policies (Chi et al., 2023; Reuss et al., 2023; Jia et al., 2025), enabling robots to acquire more complex and diverse skills (Jia et al., 2024a). Most of these methods rely on multi-modal observations for behavior learning (Zhao et al., 2023; Chi et al., 2023; Reuss et al., 2023; 2024b; Jia et al., 2024b; Reuss et al., 2024a), typically using images as state representations and language as task instructions. These approaches rely on all available modalities during inference. Due to this dependency, in scenarios of camera dropouts or noisy camera information, the performance of learned behavior can degrade (Skand et al., 2024; Hao et al., 2023b). In contrast, DisDP learns a disentangled latent space that alleviates the reliance on individual cameras during inference.

2.2 Robustness in Behavior Learning

Behavior learning suffers from generalization limitations, leading to sharp performance degradation in unfamiliar environments due to overfitting and insufficient adaptability to unseen variations (Zhang et al., 2018; Cobbe et al., 2019; Jiang et al., 2023; Kirk et al., 2023; Zare et al., 2024). To address this, various methods have been proposed to enhance robustness under modality dropout

and improve generalization in robot learning (Pari et al., 2021; Yuan et al., 2024; Liu et al., 2024b; Xie et al., 2024; Hoque et al., 2024; Becker et al., 2024). One line of work focuses on estimating missing modalities. SMIL (Ma et al., 2021) uses Bayesian meta-learning with variational inference to infer the posterior of missing inputs, while CCM (Lee et al., 2021) utilizes self-supervision to identify and discard corrupted sensor inputs before reconstructing multimodal representations. However, these methods do not account for complete sensory failure.

Several approaches have been proposed to enhance system robustness by handling missing or irrelevant modalities. Masking-based methods either drop certain modalities during training (Skand et al., 2024) or selectively mask irrelevant ones based on task relevance (Hao et al., 2023b). MIL (Hao et al., 2023a) applies masking before constructing policy representations but does not explicitly address single or multiple sensor failures. Hierarchical representation learning offers another solution. Nexus (Vasco et al., 2022) and MUSE (Vasco et al., 2021) model shared and private modality representations using dropout-based training. While Nexus aggregates modality-specific features through averaging, which limits expressiveness, MUSE employs a Product-of-Experts mechanism for more effective integration. In contrast, DisDP simplifies this process by adopting a contrastive learning framework, eliminating the need for hierarchical representations. Multi-camera setups have also been explored within reinforcement learning, using multi-view disentanglement techniques (Dunion & Albrecht, 2024) to maintain robustness when only a single camera view is available. These methods have been evaluated on Metaworld tasks and PyBullet-based environments like Panda Reach and cube grasping. DisDP extends multi-view disentanglement to imitation learning, benchmarking its performance on Colosseum and Libero, two widely used imitation learning datasets.

2.3 Multi View Disentanglement

Multi-view disentanglement has been widely explored in computer vision and multimodal learning, aiming to separate information into distinct representations for improved learning. A common approach is to decompose features into shared and private components across multiple views or modalities. Several approaches have been proposed to achieve this disentanglement. Orthogonal denoising autoencoders enforce orthogonality constraints to learn independent view representations (Ye et al., 2016). Self-supervised methods further refine this idea by explicitly minimizing inter-view overlap (Jain et al., 2023; Ke et al., 2023). These methods propose self-supervised multi-view disentanglement techniques that extract view-specific representations while preserving essential information. Their models introduce multiple loss functions to enforce alignment for shared features: orthogonalization and reconstruction-based regularization. Other methods leverage information-theoretic principles to enhance disentanglement, like FactorCL (Liang et al., 2023). It factorizes task-relevant information into shared and unique components while optimizing mutual information bounds, improving generalization by maximizing relevant features and suppressing irrelevant ones.

In DisDP, disentanglement techniques discussed above are extended to the **multi-task IL** setting, specifically within **diffusion policy frameworks**. The approach is designed to handle complex robotic manipulation tasks, with experiments conducted on diverse benchmarks. The experiments evaluate effectiveness under various sensor conditions.

3 Method

In this work, we focus on improving robustness in multi-task IL with multiple input modalities. The robot is trained to learn manipulation skills on a diverse set of tasks by imitating expert demonstrations captured with multiple cameras. These modalities can become unreliable or unavailable during deployment due to occlusion, sensor failure, or noise. The goal of this paper is to develop a robust framework capable of handling partial or imperfect inputs when such sensor issues occur.

3.1 Problem Formulation

IL aims to train an agent to perform tasks by learning from expert demonstrations. Given a dataset of expert trajectories $\mathcal{D}_{\boldsymbol{\tau}} = \{\boldsymbol{\tau}_i\}_{i=1}^{N}$, where each trajectory

$$\boldsymbol{\tau}_i = ((\boldsymbol{s}_1, \boldsymbol{a}_1), (\boldsymbol{s}_2, \boldsymbol{a}_2), \ldots, (\boldsymbol{s}_K, \boldsymbol{a}_K)) \tag{1}$$

represents a sequence of observed state-action pairs. The objective is to learn a policy $\pi(\boldsymbol{a}|\boldsymbol{s})$ that maps observations $\boldsymbol{s}$ to actions $\boldsymbol{a}$ while minimizing a some distance or divergence to the observed behavior $\mathcal{L}\left(\pi(\boldsymbol{a}|\boldsymbol{s}_k), \boldsymbol{a}_k\right)$. The exact definition of the loss $\mathcal{L}$ depends on the particular IL approach. In a multi-modal IL setting the state information contains multiple modalities, typically across different sensors. In this work these modalities include:

Language instructions $\boldsymbol{L}_k$, which provides high-level, natural language annotations for each task demonstration. Task annotations are usually provided per demonstration. However, the instruction can be simply reused at each timestep of the given trajectory $\boldsymbol{s}_k \in \boldsymbol{\tau}_i \Rightarrow \boldsymbol{L}_k \coloneqq \boldsymbol{L}_i$.

RGB images from multiple cameras $\boldsymbol{I}_k$, which capture visual scene information from different viewpoints $\boldsymbol{I}_k = (I_k^{(1)}, I_k^{(2)}, \ldots, I_k^{(C)})$

We additionally define **reliability masks** for each camera input to formulate various noise levels and the availability of different sensors $\boldsymbol{M}_k = (M_k^{(1)}, M_k^{(2)}, \ldots, M_k^{(C)})$. $M_k = \mathbf{1}$ represents a fully reliable sensor, values between 0 and 1 denote varying degrees of noise, and $M_k = \mathbf{0}$ indicates and unavailable camera input. Thus, each state in the framework is defined as

$$\boldsymbol{s}_k = (\boldsymbol{L}_k, \boldsymbol{I}_k \odot \boldsymbol{M}_k) \in \boldsymbol{\mathcal{S}}, \tag{2}$$

with $\odot$ denoting the Hadamard Product and $\boldsymbol{\mathcal{S}}$ denoting the overall state space. During the training this masking is fixed to $M_k = \mathbf{1}$. During inference, however, is masking is used to introduce noise or complete modality dropout, depending on the evaluation. Furthermore, the behavior is generally not conditioned directly on the raw sensor input but rather a learned feature embedding $\boldsymbol{z}_k = \boldsymbol{\phi}(\boldsymbol{s}_k)$, where $\boldsymbol{\phi}$ represents an encoder generating the embedding from the sensor input. Current approaches either learn a single embedding that encodes all modalities at once (Mandlekar et al., 2021b) or individual embedding per sensor input and modality (Reuss et al., 2024b; Jia et al., 2024a; Reuss et al., 2023; 2024a; Jia et al., 2025),

$$\boldsymbol{z}_k = \boldsymbol{\phi}\left(\boldsymbol{L}_k, I_k^{(1)} \odot M_k^{(1)}, I_k^{(2)} \odot M_k^{(2)}, \ldots, I_k^{(C)} \odot M_k^{(C)}\right) \text{ or} \tag{3}$$

$$\left(\boldsymbol{z}_{\boldsymbol{L},k}, \boldsymbol{z}_{I,k}^{(1)}, \boldsymbol{z}_{I,k}^{(2)}, \ldots, \boldsymbol{z}_{I,k}^{(C)}\right) = \boldsymbol{\phi}\left(\boldsymbol{L}_k, I_k^{(1)} \odot M_k^{(1)}, I_k^{(2)} \odot M_k^{(2)}, \ldots, I_k^{(C)} \odot M_k^{(C)}\right). \tag{4}$$

Theoretically, learning individual embeddings provides mechanisms to improve robustness against modality dropout. In practice, however, the learned policies usually still require the presence of all embeddings and assume reliable information for each. This work in contrast, does not learn embeddings for individual sensors nor single embeddings across all sensors but instead explicitly learns shared embeddings $\boldsymbol{v}$ across sensors and private embeddings $\boldsymbol{u}$ for each sensor

$$\left(\boldsymbol{z}_{\boldsymbol{L},k}, \boldsymbol{z}_{I,k}^{(1)}, \boldsymbol{z}_{I,k}^{(2)}, \ldots, \boldsymbol{z}_{I,k}^{(C)}\right) \Rightarrow \left(\boldsymbol{z}_{\boldsymbol{L},k}, \left(\boldsymbol{v}_{I,k}^{(1)}, \boldsymbol{u}_{I,k}^{(1)}\right), \left(\boldsymbol{v}_{I,k}^{(2)}, \boldsymbol{u}_{I,k}^{(2)}\right), \ldots, \left(\boldsymbol{v}_{I,k}^{(C)}, \boldsymbol{u}_{I,k}^{(C)}\right)\right) \tag{5}$$

The shared embeddings $\boldsymbol{v}^{(c)}$ contain information that sensor c shares with other sensors, while the private embeddings $\boldsymbol{u}^{(c)}$ contain information that is unique to the sensor. This formulation allows the policy to learn a more robust representation of unreliable sensors. If the sensor c drops out, the private information $\boldsymbol{u}^{(c)}$ of the sensor is not available, however, the information that would have been contained in the shared embedding $\boldsymbol{v}^{(c)}$ is covered by the other sensors.

Finally, recent work introducing action chunking(Zhao et al., 2023) has shown that predicting a sequence of actions generally performs better than generating single step actions. Following this insight the action space is redefined as

$$\bar{\boldsymbol{a}}_k = (\boldsymbol{a}_k, \boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_{k+H}) \in \boldsymbol{\mathcal{A}}^H, \tag{6}$$

where H is the prediction horizon, $\boldsymbol{\mathcal{A}}$ denotes the action space and the sequence of actions $(\boldsymbol{a}_k, \boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_{k+H})$ was observed in any of the demonstrated trajectories $\boldsymbol{\tau}_i$.

The final policy is represented as $\bar{\boldsymbol{a}}_k \sim \pi\left(\bar{\boldsymbol{a}}_k | \boldsymbol{\phi}(\boldsymbol{s}_k)\right)$ and trained using the dataset

$$\mathcal{D} = \bigcup_{\boldsymbol{\tau} \in \mathcal{D}_{\boldsymbol{\tau}}} \left\{ (\bar{\boldsymbol{a}}, \boldsymbol{s}) | (\bar{\boldsymbol{a}}, \boldsymbol{s}) \in \boldsymbol{\tau} \right\}, \tag{7}$$

which contains pairs of action sequences and states across all demonstrated trajectories. Here, the union $\bigcup$ allows for potentially duplicate entries in the final dataset to maintain the statistical occurrence of state-action pairs.

3.2 Preliminaries

Score-based Diffusion Policy: We model the policy using a continuous-time denoising diffusion process based on the EDM formulation(Karras et al., 2022; Lu & Song, 2024). Denoising diffusion models aim to reverse a stochastic noising process that transforms data into Gaussian noise (Song et al., 2020), enabling the generation of new samples that match the data distribution. The denoising process follows a stochastic differential equation (SDE):

$$\mathrm{d}\bar{\boldsymbol{a}} = \left(\beta_t \sigma_t - \dot{\sigma}_t\right) \sigma_t \nabla_{\bar{\boldsymbol{a}}} \log p_t(\bar{\boldsymbol{a}} | \boldsymbol{\phi}(\boldsymbol{s})) \mathrm{d}t + \sqrt{2\beta_t} \sigma_t \mathrm{d}B_t, \tag{8}$$

where β_t controls the noise injection, B_t is a standard Wiener process, and $\nabla_{\bar{\boldsymbol{a}}} \log p_t(\bar{\boldsymbol{a}} | \boldsymbol{\phi}(\boldsymbol{s}))$ is the score function that guides samples toward regions of high data density. A neural network approximates the score $\nabla_{\bar{\boldsymbol{a}}} \log p_t(\bar{\boldsymbol{a}} | \boldsymbol{\phi}(\boldsymbol{s}))$ by minimizing the Score Matching objective (Vincent, 2011),

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{\sigma_t, \bar{\boldsymbol{a}}, \boldsymbol{s}, \boldsymbol{\epsilon}} \left[\alpha(\sigma_t) \| F_\theta(\bar{\boldsymbol{a}} + \boldsymbol{\epsilon}, \boldsymbol{\phi}(\boldsymbol{s}), \sigma_t) - \bar{\boldsymbol{a}} \|_2^2 \right], \tag{9}$$

where $F_\theta(\bar{\boldsymbol{a}} + \boldsymbol{\epsilon}, \boldsymbol{\phi}(\boldsymbol{s}), \sigma_t)$ is the trainable network. During training, noise sampled from a predefined distribution is added to an action sequence, and the network learns to predict the denoised actions. After training, new action sequences are generated by sampling an action $\bar{\boldsymbol{a}}_T \sim \mathcal{N}(0, \sigma_T^2 I)$ from the prior distribution and progressively denoising it by approximating the reverse SDE. The efficient **DDIM ODE-Solver** (Song et al., 2020) enables denoising in just a few steps (Reuss et al., 2023).

Multi-view disentanglement: This technique separates representations into shared, $\boldsymbol{v}$, and private, $\boldsymbol{u}$, components across sensors and modalities. Shared representations capture the global information consistent across multiple views, while private representations encode unique information that is specific to a camera view. Separating the information of multiple sensors into shared and private components enhances the robustness as shared information provides stability in scenarios where information from a certain modality is either absent or unreliable. Meanwhile, private representations embed fine-grained details which can improve the task performance when available.

3.3 Disentangled Diffusion Policy

Disentangled Diffusion Policy (DisDP) combines a Transformer based encoder-decoder diffusion model (Reuss et al., 2023; 2024b) with multi-view disentanglement, as illustrated in Figure 2. A more general architecture for any given IL method integrating disentanglement can be found in Appendix A. In the first step, every camera input $I_k^{(c)}$ is embedded using a separate vision encoder. These vision-embeddings are processed through disentanglement branches to obtain a shared embedding $\boldsymbol{v}_k^{(c)}$ and a private embedding $\boldsymbol{u}_k^{(c)}$.

The shared embedding module extracts global features, that are consistent across all camera views $I_k^{(1:C)}$. By focusing on features that remain stable across viewpoints, the shared-embedding encoder provides a robust foundation for downstream tasks, especially when one or more cameras become unreliable, occluded, or noisy. The private embedding module captures fine-grained and

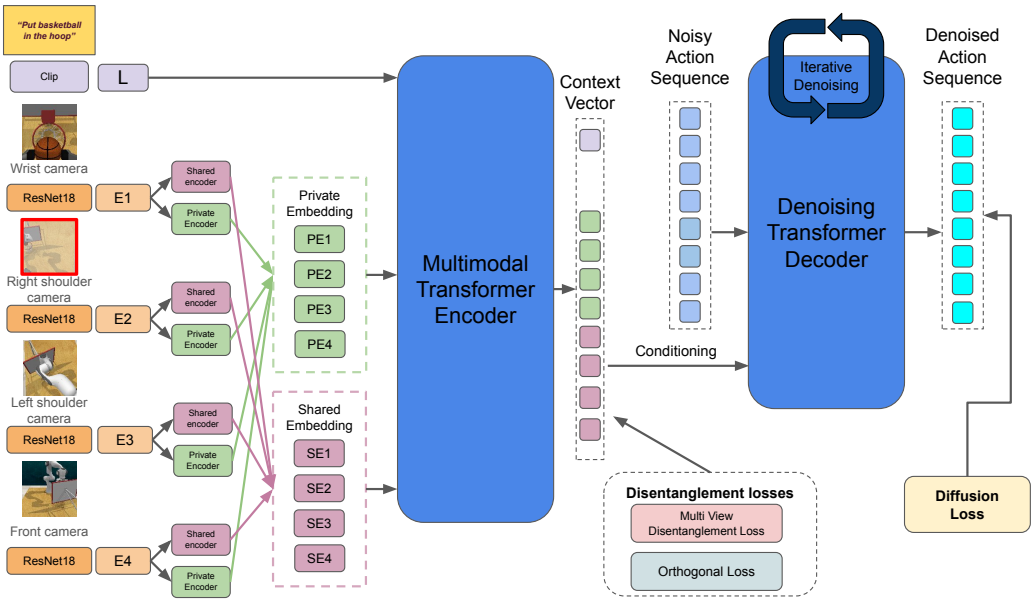

Figure 2: Overview of Disentangled Diffusion Policy (DisDP). The model processes multi-view image inputs by separating them into shared and private representations. Each camera input is encoded using ResNet-18, followed by disentanglement modules that extract shared embeddings across all views and private embeddings for individual views. These embeddings are processed by a multimodal transformer encoder and serve as conditioning inputs to the denoising transformer decoder for action prediction. The model is trained with a combination of diffusion loss, multi-view disentanglement loss, and orthogonality loss to enforce representation separation. This structured representation learning enhances robustness to sensor noise, failures, and environmental variations.

view-specific details for each camera view $I_k^{(c)}$. These private features enrich the policy with information unique to each perspective, preserving distinctive cues when global signals are insufficient.

The effective separation of shared and private features is ensured using a contrastive learning approach based on the $\mathrm{InfoNCE}(\boldsymbol{x}, \boldsymbol{x}_+, \boldsymbol{x}_-)$ loss (Oord et al., 2018; Chen et al., 2020). The contrastive learning loss requires positive $\boldsymbol{x}_+$ and negative samples $\boldsymbol{x}_-$ for each point $\boldsymbol{x}$. The $\mathrm{InfoNCE}$ loss then rewards embeddings that are close to positive samples while punishing embeddings that are close to negative samples.

For the shared embedding $\boldsymbol{v}^c$ of sensor c we obtain the positive samples $\boldsymbol{v}_+^c$ by sampling shared embeddings of different sensors at the same state. While negative samples $\boldsymbol{v}_-^c$ are sampled from shared embeddings of different states. The corresponding disentanglement loss is defined as

$$\mathcal{L}_{\text{shared}} = \mathbb{E}_{\boldsymbol{s}\in\mathcal{D}, c\in C, \boldsymbol{v}^c\in\boldsymbol{\phi}(\boldsymbol{s})}\, \mathrm{InfoNCE}(\boldsymbol{v}^c, \boldsymbol{v}_+^c, \boldsymbol{v}_-^c). \tag{10}$$

For the private embedding $\boldsymbol{u}^c$ of sensor c the positive samples $\boldsymbol{u}_+^c$ are drawn form the same camera at different states and the negative samples $\boldsymbol{u}_-^c$ are drawn from any other sensor at any state. The corresponding disentanglement loss is defined analogously to the shared loss

$$\mathcal{L}_{\text{private}} = \mathbb{E}_{\boldsymbol{s}\in\mathcal{D}, c\in C, \boldsymbol{u}^c\in\boldsymbol{\phi}(\boldsymbol{s})}\, \mathrm{InfoNCE}(\boldsymbol{u}^c, \boldsymbol{u}_+^c, \boldsymbol{u}_-^c). \tag{11}$$

Both loss functions can be combined into the disentanglement loss

$$\mathcal{L}_{\text{disent}} = \mathcal{L}_{\text{shared}} + \mathcal{L}_{\text{private}}, \tag{12}$$

which ensure maximization of similarity among the shared representation, minimization of similarity between shared and private representations and minimization of similarity between individual

private representations. Apart from the contrastive objective, DisDP adds an orthogonality loss

$$\mathcal{L}_{\text{ortho}} = \mathbb{E}_{\boldsymbol{s}\in\mathcal{D}, c\in C, (\boldsymbol{v}^c, \boldsymbol{u}^c), \in \boldsymbol{\phi}(\boldsymbol{s})} \langle \boldsymbol{v}^c, \boldsymbol{u}^c \rangle^2, \tag{13}$$

to further disentangle the shared and private embeddings by minimizing the squared dot product $\langle \cdot, \cdot \rangle$ between them for each camera. Together with the diffusion loss, this results in the final loss

$$\mathcal{L} = \mathcal{L}_{\text{diffusion}} + \lambda_{\text{disent}} \cdot \mathcal{L}_{\text{distent}} + \lambda_{\text{ortho}} \cdot \mathcal{L}_{\text{ortho}}, \tag{14}$$

where λ_{disent} and λ_{ortho} are hyperparameters scaling the importance of the disentanglement and orthogonality loss.

4 Evaluation

The experiments conducted in this paper try to answer 4 research questions increasing in difficulty towards robustness and 1 research question with focus on interpretability:
RQ1: Does disentanglement affect the performance of IL policies?
RQ2: Do disentangled latent spaces improve resilience to noisy sensor input?
RQ3: Do disentangled latent spaces improve resilience to complete sensor dropout?
RQ4: How resilient are policies to environmental perturbations and sensor dropout?
RQ5: Does disentanglement results in more interpretable latent spaces?

To answer these questions all policies are evaluated on two sota IL benchmarks, The Colosseum (Pumacay et al., 2024) and Libero (Liu et al., 2024a). Both environments provide multi-camera image observations with 5 cameras for The Colosseum and 2 cameras for Libero. The Colosseum is constructed using tasks from RLBench (James et al., 2020), to benchmark complex robot manipulation tasks. It has 20 tabletop tasks with different variations in each task, including changes in lighting, texture, object colors and properties. Libero consists of diverse robot manipulation tasks categorized into object, spatial, goal, and long-horizon tasks. These tasks evaluate robotic skills on different skill ranges, making it a comprehensive benchmark for generalization in robotic manipulation. In both benchmarks, Policy performance is assessed using success rate, defined as the percentage of rollouts that successfully complete the task within a specified number of steps.

4.1 Evaluated approaches

During the evaluation, three baselines are considered:
BC: Behavior Cloning (BC) is usually used as a default baseline for imitation learning. We apply an encoder-decoder Transformer architecture to perform action prediction, which is optimized by Mean Squared Error (MSE).
BESO-ACT: BEhavior generation with ScOre-based Diffusion Policies (BESO) (Reuss et al., 2023) is a diffusion-based policy that represents the denoising process using a continuous Stochastic-Differential Equation (SDE). Beyond that, we build BESO-ACT by using the same Transformer in BC and applying action chunking (Zhao et al., 2023).
BESO-ACT-dropout: This baseline uses BESO-ACT but introduces random modality dropout in training at a rate of 10 percent to gain robustness.

Our Contributed methods are:
DisBC: DisBC extends the BC baseline by introducing disentangled latent spaces.
DisDP: DisDP integrates disentangled representations in the BESO-ACT architecture

4.2 Experimental Setup

The Colosseum: With regards to **RQ1**, experiments are conducted on 10 of the 20 Colosseum tasks: basketball in hoop, close box, close laptop lid, hockey, meat on grill, move hanger, open drawer, reach and drag, scoop with spatula, and slide block to target. These tasks were selected

based on their strong performance using the baseline method, ensuring a fair comparison. The proposed methods and baselines are trained on the *no-variation* setting within the Colosseum suite for 200 epochs on the same hyperparameters to avoid biases. The trained models are evaluated on noisy camera sensor input and complete dropout to address **RQ2** and **RQ3**. Included cameras are: **0** left view, **1** right view, **2** wrist view and **3** front view. The bird view is disregarded because initial performance did not increase when including it. Dual camera dropouts are only reported for **0 1** and **1 2** because other combinations achieve low success rate for all methods. Regarding **RQ4**, the trained models are evaluated on 8 different Colosseum variations: no-variation, background texture, camera pose, distractor, light color, object color, table color, table texture. The dataset contains 100 demonstrations for each task with images captured from the five camera views. The policies are evaluated using three seeds, with 25 rollouts per task and a maximum of 300 steps per rollout.
Libero: Addressing **RQ1** policies are evaluated on 3 of the 4 categories, excluding long-horizon tasks for computation reasons. Models are trained for 50 epochs on 60 percent of demonstrations on same hyperparameters to ensure fair comparison. Libero includes 2 camera views: Agent camera **0** and in-hand camera **1**. Regarding **RQ3**, policies are evaluated on dropping out either the agent or in-hand view. The methods are evaluated using three different seeds with 25 rollouts per task in each dataset split. Each episode has maximum 260 steps per rollout.

4.3 Result analysis

The following section discusses the 5 introduced research questions with regard to the experimental results on The Colosseum (Pumacay et al., 2024) and Libero (Liu et al., 2024a) benchmarks. Further results regarding evaluations can be found in Appendix B.

Benchmark	BC	DisBC	BESO-ACT	BESO-ACT-Dropout	DisDP
Colosseum	0.361 ± 0.11	0.540 ± 0.08	0.709 ± 0.03	0.435 ± 0.04	$\mathbf{0.896 \pm 0.05}$
Libero - Object	0.684 ± 0.00	0.736 ± 0.02	0.752 ± 0.00	0.514 ± 0.05	$\mathbf{0.816 \pm 0.02}$
Libero - Spatial	± 0.00	0.583 ± 0.02	0.580 ± 0.03	0.552 ± 0.04	$\mathbf{0.701 \pm 0.04}$
Libero - Goal	$0.047 \pm$	$0.065 \pm$	0.576 ± 0.02	0.418 ± 0.05	$\mathbf{0.680 \pm 0.09}$

Table 1: **The Colosseum and Libero results for all policies with reliable modality inputs.** Using disentangled IL policies does overall improve performance on both benchmarks.

RQ1: Does disentanglement affect the performance of IL policies?

The first research question aims at analyzing the quality of policies when adding disentanglement, because of the trade-off between performance and interpretability. Table 1 displays the results for all three baselines and the two proposed methods using disentangled shared and private embeddings. In both benchmarks, using the disentangled version of the baseline does improve overall performance. DisDP achieves 0.896 success rate compared to the 0.709 of BESO-ACT on the Colosseum tasks. It also improves results on the Libero benchmark between 0.06 and 0.12, compared to the BESO-ACT baseline. In general, disentangled IL policies do improve overall performance.

RQ2: Do disentangled latent spaces improve resilience to noisy sensor input?

The first step of unreliable sensors include adding noise to the camera input of the model, as this is a common failure case. For all methods, the overall performance on The Colosseum benchmark does drop significantly, as shown in Table 2. The disentangled methods still outperform their corresponding baseline methods and with less performance loss. Especially the DisBC still performs similar to the non-noisy results. It even outperforms the DisDP, when dropping out the front view **3** and both dual combinations **0 1** and **1 2**. The traditional BC, on the other hand, completely fails when confronted with noisy sensor inputs. In the noisy scenario, the BESO-ACT-dropout is also able to retain more of its original performance, compared to the BESO-ACT. These observations answer **RQ2**: Disentangled IL policies are much more resilient towards noisy sensor inputs.

RQ3: Do disentangled latent spaces improve resilience to complete sensor dropout?

Noisy	BC	DisBC	BESO-ACT	BESO-ACT-Dropout	DisDP
None	0.361 ± 0.11	0.540 ± 0.08	0.709 ± 0.03	0.435 ± 0.04	$\mathbf{0.896 \pm 0.05}$
0	0.160 ± 0.05	0.444 ± 0.04	0.000 ± 0.00	0.020 ± 0.02	$\mathbf{0.568 \pm 0.11}$
1	0.028 ± 0.03	0.496 ± 0.05	0.288 ± 0.07	0.326 ± 0.07	$\mathbf{0.500 \pm 0.12}$
2	0.100 ± 0.02	0.196 ± 0.03	0.008 ± 0.01	0.280 ± 0.01	$\mathbf{0.306 \pm 0.08}$
3	0.130 ± 0.02	$\mathbf{0.440 \pm 0.02}$	0.252 ± 0.03	0.210 ± 0.07	0.280 ± 0.04
0 1	0.020 ± 0.01	$\mathbf{0.420 \pm 0.01}$	0.000 ± 0.00	0.020 ± 0.01	0.378 ± 0.05
1 2	0.080 ± 0.07	$\mathbf{0.370 \pm 0.02}$	0.000 ± 0.00	0.186 ± 0.04	0.172 ± 0.04

Table 2: **Colosseum no variation Dataset Evaluation with Noisy Camera Views**. The numbers in the column *Noisy* correspond to the specific camera: **0** left view, **1** right view, **2** wrist view, and **3** front view. The evaluation examines how noisy sensors affect task success rates and assesses the resilience of different methods under these conditions. The disentangled methods perform much better compared to their baseline implementations. Especially the DisBC has a small decrease in performance, when adding noise.

Across both Libero and Colosseum, BESO-ACT and BESO-ACT-Dropout experience significant performance drops when critical camera views are unavailable, highlighting their reliance on complete visual input. Notably, BESO-ACT-Dropout fails to mitigate sensor failures, showing that naive modality dropout during training does not improve robustness but instead leads to the loss of important task-relevant information. In Colosseum, displayed in Table 3, evaluation is conducted with four cameras, providing redundancy and robustness to sensor failures due to overlapping viewpoints. In contrast, Libero only has two cameras in total, relying heavily on both of them. Table 4 shows the impact of sensor dropout in Libero, where performance declines sharply across the object, spatial, and goal task suites when the **0** agent or **1** in-hand camera is masked.

Masked	BC	DisBC (Ours)	BESO-ACT	BESO-ACT-Dropout	DisDP (Ours)
None	0.361 ± 0.11	0.540 ± 0.08	0.709 ± 0.03	0.435 ± 0.04	$\mathbf{0.896 \pm 0.05}$
0	0.096 ± 0.01	0.206 ± 0.03	0.068 ± 0.05	0.096 ± 0.01	$\mathbf{0.440 \pm 0.03}$
1	0.120 ± 0.02	0.140 ± 0.03	0.196 ± 0.04	0.168 ± 0.02	$\mathbf{0.632 \pm 0.04}$
2	0.048 ± 0.01	0.228 ± 0.01	0.292 ± 0.03	0.100 ± 0.03	$\mathbf{0.420 \pm 0.02}$
3	0.028 ± 0.02	$\mathbf{0.096 \pm 0.01}$	0.040 ± 0.03	0.004 ± 0.00	0.060 ± 0.03
0 1	0.056 ± 0.01	0.100 ± 0.02	0.028 ± 0.02	0.048 ± 0.01	$\mathbf{0.196 \pm 0.05}$
1 2	0.000 ± 0.00	0.092 ± 0.01	0.070 ± 0.01	0.040 ± 0.02	$\mathbf{0.192 \pm 0.07}$

Table 3: **Colosseum no variation Dataset Evaluation with Masked Camera Views**. Columns 0, 1, 2, and 3 represent the masking of individual cameras: 0 (left view), 1 (right view), 2 (wrist view), and 3 (front view). The evaluation examines how sensor failures affect task success rates and assesses the resilience of different methods under these conditions.

Unlike BC and BESO-ACT, the disentangled-based methods DisBC and DisDP exhibit greater resilience to sensor failures, maintaining higher success rates across masked conditions. DisDP consistently outperforms all baselines, even when multiple modalities are missing. For instance, in Colosseum when **0** left view is masked, DisDP retains a success rate of 0.440, significantly higher than BC with 0.096 and BESO-ACT with 0.068, as shown in Table 3. These results highlight DisDP's ability to adapt to missing sensory inputs through its disentanglement-based structure. Additionally, the front view emerges as the most critical modality in Colosseum, where its removal leads to the largest performance drop. A similar trend is observed in Libero, where DisDP maintains the highest performance for all task suites and masked cameras, except for Goal - 0, where BESO-ACT achieves the highest result. This observation further confirms the reliance on full visual input

Masked	BC	DisBC	BESO-ACT	BESO-ACT-Dropout	DisDP
Object - None	0.684 ± 0.00	0.736 ± 0.02	0.752 ± 0.00	0.514 ± 0.05	$\mathbf{0.816 \pm 0.02}$
Spatial - None	0.556 ± 0.00	0.583 ± 0.02	0.580 ± 0.03	0.552 ± 0.04	$\mathbf{0.701 \pm 0.04}$
Goal - None	-	-	0.576 ± 0.02	0.418 ± 0.05	$\mathbf{0.680 \pm 0.09}$
Object - 0	0.000 ± 0.00	0.110 ± 0.03	0.204 ± 0.00	0.004 ± 0.00	$\mathbf{0.295 \pm 0.04}$
Spatial - 0	0.000 ± 0.00	0.000 ± 0.00	0.028 ± 0.00	0.023 ± 0.00	$\mathbf{0.144 \pm 0.02}$
Goal - 0	-	-	$\mathbf{0.084 \pm 0.01}$	0.040 ± 0.00	0.004 ± 0.00
Object - 1	0.000 ± 0.00	0.000 ± 0.00	0.012 ± 0.01	0.000 ± 0.00	$\mathbf{0.226 \pm 0.03}$
Spatial - 1	0.000 ± 0.00	0.004 ± 0.00	0.004 ± 0.00	0.023 ± 0.04	$\mathbf{0.112 \pm 0.00}$
Goal - 1	-	-	0.012 ± 0.00	0.004 ± 0.00	$\mathbf{0.200 \pm 0.04}$

Table 4: **Libero dataset evaluation**: The evaluation examines three task suites—Object, Spatial, and Goal—across three conditions: normal (all cameras available), agent view camera masked (0), and in-hand camera masked (1). The results demonstrate the effect of modality dropout on task success and highlight that policies trained with disentangled methods exhibit better adaptability to missing sensory inputs.

of BESO-ACT. The impact of masking the **1** in-hand camera is even more severe, with DisDP still achieving 0.113, whereas BESO-ACT drops to 0.009.

DisDP achieves the highest performance retention under modality dropout, demonstrating that disentangled representations effectively preserve task-relevant features despite missing inputs. While DisBC also leverages shared and private representation separation and improves robustness over BC, it does not match the adaptability of DisDP, which benefits from diffusion policies in addition to disentangled representations. This analysis answers **RQ3** and shows that disentangled representations help retain important task information and handle sensor dropout. Separating shared and private representations makes models more adaptable to sensor failures, reducing dependence on a single modality and improving robustness in different conditions.

RQ4: How resilient are policies to environmental perturbations and sensor dropout?

To evaluate the robustness of policies on environmental perturbations and modality dropouts, Colosseum provides 7 different variations for all tasks. Previous experiments showed that the diffusion-based methods perform best on The Colosseum, therefore only those two methods are evaluated on the variations of The Colosseum.

Figure 3 presents the evaluations on the environmental variations from The Colosseum. The no-variation condition serves as the baseline, achieving the highest performance across all tasks. Figures for individual tasks are given in Appendix B. The results indicate the degradation in performance on environmental perturbations. The performance decreases further when certain camera views are dropped out. Overall, DisDP demonstrates greater robustness compared to BESO-ACT, particularly in handling object color, table color, and background texture variations.

The results show that disentangled representations help handle environmental changes. Even without training on these variations, our method remained more robust. Variations in camera pose, table texture, and lighting were the most disruptive factors, affecting the model's spatial reasoning and fine-grained perception. These observations address **RQ4**, as it highlights how disentangled representations enable improved generalization to environmental perturbations by preserving task-relevant features while filtering out irrelevant variations.

RQ5: Does disentanglement results in more interpretable latent spaces?

DisDP's superior performance in Libero and Colosseum shows that separating shared and private components enhances robustness and adaptability under both normal and unreliable camera conditions. To investigate the interpretability of the disentangled latent space, we examine the saliency maps of the learned shared and private representations in Figure 4, using the close-box task as an

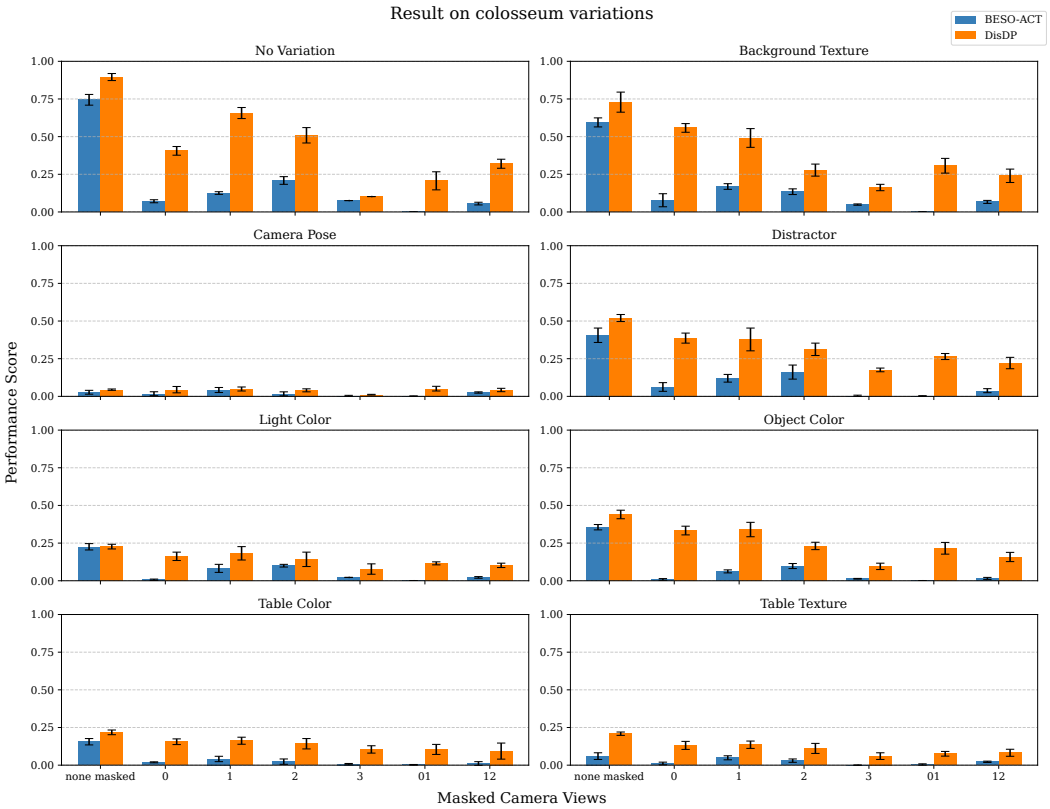

Figure 3: **Colosseum results on variations, comparison between BESO-ACT and DisDP**. The no-variation condition serves as a baseline, showing the highest performance. Spatial, textural, and lighting variations significantly impact success rates, with camera pose and table texture masking causing the most degradation.

example. The shared representation focuses on box edges, a crucial feature for proper alignment and closure, ensuring that essential task information remains consistent across views. This cross-view consistency allows the model to retain key information, even when some camera inputs are missing or degraded. In contrast, the private representations capture view-specific details, such as robot joints and table shadows, which provide additional contextual information for precise manipulation.

By clearly separating shared and private representations, disentanglement improves generalization and enhances robustness to sensor failures. When a camera is not available, the shared representation still preserves essential task features, enabling stable execution. In addition, disentanglement enhances interpretability by clarifying the distinct role of each view in decision-making. To further analyze the structure of disentangled representations, we used Uniform Manifold Approximation and Projection (UMAP) to plot the shared and private embeddings, displayed in Figure 5. This separation confirms that disentanglement structures the latent space in a way that improves both interpretability and robustness.

5 Conclusion

This work introduced Disentangled Diffusion Policy (DisDP), a method for improving robustness in IL by leveraging multi-view disentanglement. By structuring sensor inputs into shared and private representations, DisDP enhances the model's ability to handle sensor noise, dropouts and environmental variations better. Our evalua-

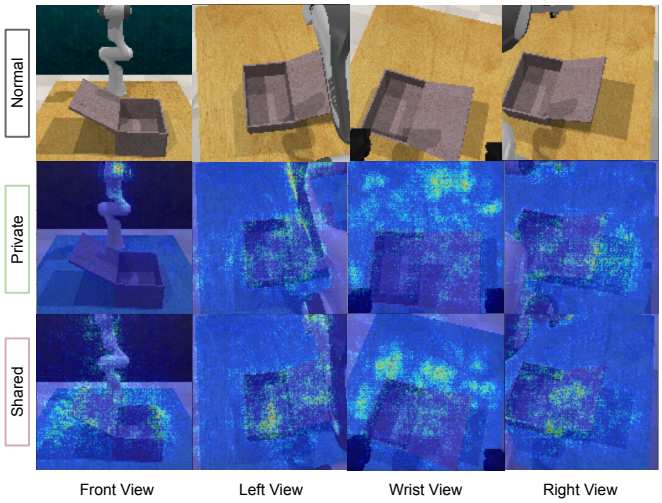

Figure 4: **Saliency maps for disentangled embeddings**. In the close box task, the shared embeddings capture the box edges, which are crucial for task completion and visible across different views. In contrast, the private embeddings focus on specific details, such as robot joints and table shadows, which contribute to task execution, while others capture unique but less relevant scene elements.

tions on The Colosseum and Libero demonstrate that disentangled methods achieve better performance than their baseline implementation, even when all sensors are available. Evaluations with noisy or unreliable sensors demonstrated the robustness improvement through disentangled IL methods. Furthermore, disentanglement additionally provides more robustness towards environmental changes, making models more robust in general. The separation of private and shared embeddings allows for visualization of the latent space through Gradient-weighted Class Activation Mapping (Grad-CAM) and UMAP. These visualizations give insight into the focus of the model and how private and shared embeddings are separated.

Limitations of DisDP can be observed when looking at camera dropout combinations. If specific combinations of cameras are not available, disentanglement does not retain information to complete tasks reliably. Furthermore, less cameras decrease the efficiency of disentangled methods, because the shared embedding has less overlap between viewpoints.

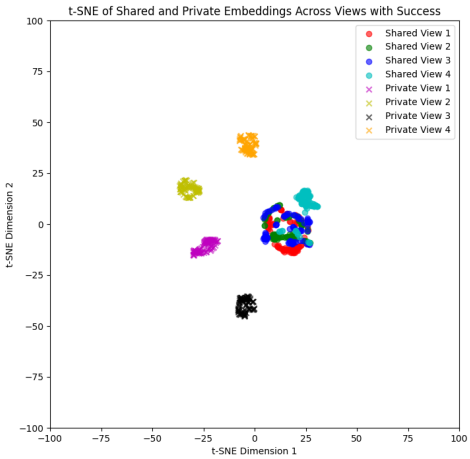

Figure 5: UMAP plot - shared private disentanglement.The visualization confirms the separation process: shared embeddings from all four cameras cluster at the center, while private embeddings remain dispersed, capturing view-specific features.

Future work will focus on improving robustness under less modality inputs and to reduce performance loss if more then one modality is not available. The next steps will also include real robot experiments, to confirm the proposed method outside of simulation. Furthermore, including other sensor modalities beside vision would be interesting and could enhance model performance.

References

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Philipp Becker, Sebastian Mossburger, Fabian Otto, and Gerhard Neumann. Combining reconstruction and contrastive methods for multimodal representations in rl. In *Reinforcement Learning Conference*, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, pp. 02783649241273668, 2023.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *International conference on machine learning*, pp. 1282–1289, California, 2019. PMLR.

Atalay Donat, Xiaogang Jia, Xi Huang, Aleksandar Taranovic, Denis Blessing, Ge Li, Hongyi Zhou, Hanyi Zhang, Rudolf Lioutikov, and Gerhard Neumann. Towards fusing point cloud and visual representations for imitation learning. *arXiv preprint arXiv:2502.12320*, 2025.

Mhairi Dunion and Stefano V Albrecht. Multi-view disentanglement for reinforcement learning with multiple cameras. *arXiv preprint arXiv:2404.14064*, 2024.

Yilun Hao, Ruinan Wang, Zhangjie Cao, Zihan Wang, Yuchen Cui, and Dorsa Sadigh. Masked imitation learning: Discovering environment-invariant modalities in multimodal demonstrations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–7, 2023a. DOI: 10.1109/IROS55552.2023.10341728.

Yilun Hao, Ruinan Wang, Zhangjie Cao, Zihan Wang, Yuchen Cui, and Dorsa Sadigh. Masked imitation learning: Discovering environment-invariant modalities in multimodal demonstrations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–7. IEEE, 2023b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ryan Hoque, Ajay Mandlekar, Caelan Garrett, Ken Goldberg, and Dieter Fox. Intervengen: Interventional data generation for robust and data-efficient robot imitation learning. *arXiv preprint arXiv:2405.01472*, 2024.

Nihal Jain, Praneetha Vaddamanu, Paridhi Maheshwari, Vishwa Vinay, and Kuldeep Kulkarni. Self-supervised multi-view disentanglement for expansion of visual collections. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 841–849, 2023.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Xiaogang Jia, Denis Blessing, Xinkai Jiang, Moritz Reuss, Atalay Donat, Rudolf Lioutikov, and Gerhard Neumann. Towards diverse behaviors: A benchmark for imitation learning with human demonstrations. *arXiv preprint arXiv:2402.14606*, 2024a.

Xiaogang Jia, Qian Wang, Atalay Donat, Bowen Xing, Ge Li, Hongyi Zhou, Onur Celik, Denis Blessing, Rudolf Lioutikov, and Gerhard Neumann. Mail: Improving imitation learning with mamba. *arXiv preprint arXiv:2406.08234*, 2024b.

Xiaogang Jia, Atalay Donat, Xi Huang, Xuan Zhao, Denis Blessing, Hongyi Zhou, Hanyi Zhang, Han A Wang, Qian Wang, Rudolf Lioutikov, et al. X-il: Exploring the design space of imitation learning policies. *arXiv preprint arXiv:2502.12330*, 2025.

Yiding Jiang, J. Zico Kolter, and Roberta Raileanu. On the importance of exploration for generalization in reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 12951–12986, New Orleans, USA, 2023. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2a4310c4fd24bd336aa2f64f93cb5d39-Paper-Conference.pdf.

Joshua Jones, Oier Mees, Carmelo Sferrazza, Kyle Stachowicz, Pieter Abbeel, and Sergey Levine. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. *arXiv preprint arXiv:2501.04693*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Guanzhou Ke, Yang Yu, Guoqing Chao, Xiaoli Wang, Chenyang Xu, and Shengfeng He. Disentangling multi-view representations beyond inductive bias. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2582–2590, 2023.

Robert Kirk, Amy Zhang, Edward Grefenstette, and Tim Rocktäschel. A survey of zero-shot generalisation in deep reinforcement learning. *Journal of Artificial Intelligence Research*, 76:201–264, 2023.

Michelle A Lee, Matthew Tan, Yuke Zhu, and Jeannette Bohg. Detect, reject, correct: Crossmodal compensation of corrupted sensors. In *2021 IEEE international conference on robotics and automation (ICRA)*, pp. 909–916. IEEE, 2021.

Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. *arXiv preprint arXiv:2212.03858*, 2022.

Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36, 2024a.

Rui Liu, Amisha Bhaskar, and Pratap Tokekar. Adaptive visual imitation learning for robotic assisted feeding across varied bowl configurations and food types. *arXiv preprint arXiv:2403.12891*, 2024b.

Zeyi Liu, Cheng Chi, Eric Cousineau, Naveen Kuppuswamy, Benjamin Burchfiel, and Shuran Song. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. In *8th Annual Conference on Robot Learning*, 2024c.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *5th Annual Conference on Robot Learning*, 2021a. URL https://openreview.net/forum?id=JrsfBJtDFdI.

Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2021b.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends® in Robotics*, 7(1-2): 1–179, 2018.

Jyothish Pari, Nur Muhammad Shafiullah, Sridhar Pandian Arunachalam, and Lerrel Pinto. The surprising effectiveness of representation learning for visual imitation. *arXiv preprint arXiv:2112.01511*, 2021.

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191*, 2024.

Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. In *Robotics: Science and Systems*, 2023.

Moritz Reuss, Jyothish Pari, Pulkit Agrawal, and Rudolf Lioutikov. Efficient diffusion transformer policies with mixture of expert denoisers for multitask learning. *arXiv preprint arXiv:2412.12953*, 2024a.

Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. In *Robotics: Science and Systems*, 2024b.

Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Skand Skand, Bikram Pandit, Chanho Kim, Li Fuxin, and Stefan Lee. Simple masked training strategies yield control policies that are robust to sensor failure. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=AsbyZRdqPv.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. URL https://arxiv.org/abs/2010.02502.

Abitha Thankaraj and Lerrel Pinto. That sounds right: Auditory self-supervision for dynamic robot manipulation. In *Conference on Robot Learning*, pp. 1036–1049. PMLR, 2023.

Miguel Vasco, Hang Yin, Francisco S Melo, and Ana Paiva. How to sense the world: Leveraging hierarchy in multimodal perception for robust reinforcement learning agents. *arXiv preprint arXiv:2110.03608*, 2021.

Miguel Vasco, Hang Yin, Francisco S Melo, and Ana Paiva. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255, 2022.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. DOI: 10.1162/NECO_a_00142.

Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3153–3160. IEEE, 2024.

TengQi Ye, Tianchun Wang, Kevin McGuinness, Yu Guo, and Cathal Gurrin. Learning multiple views with orthogonal denoising autoencoders. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part I 22*, pp. 313–324. Springer, 2016.

Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. *arXiv preprint arXiv:2407.15815*, 2024.

Maryam Zare, Parham M. Kebria, Abbas Khosravi, and Saeid Nahavandi. A survey of imitation learning: Algorithms, recent developments, and challenges. *IEEE Transactions on Cybernetics*, 54(12):7173–7186, 2024. DOI: 10.1109/TCYB.2024.3395626.

Amy Zhang, Nicolas Ballas, and Joelle Pineau. A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*, 1(1), 2018.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

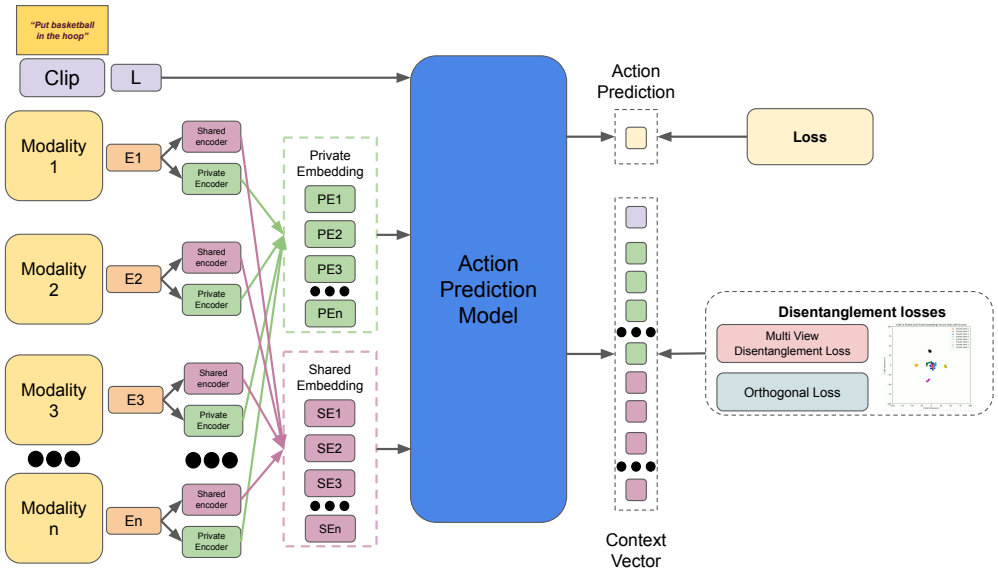

Figure 6: General Architecture

Supplementary Materials

The following content was not necessarily subject to peer review.

A General Architecture for Disentanglement in Imitation Learning

A general architecture for introducing disentangled latent spaces in Imitation learning is shown in Figure 6 with n number of modalities.

B Evaluation on environment variations

We have added a few individual task results for the environment variations for the Colosseum benchmark to analyze the effect of perturbations on tasks.

Simple tasks like close box remain relatively robust across most of variations except the camera pose. The performance of open drawer task reduces a lot under variations like table color, texture, and even background texture and lighting changes. It shows the representations are still focusing on table pixels. Tasks like reach and drag, and basketball experience reduced performance by adding distractor objects, which confuses the policies.

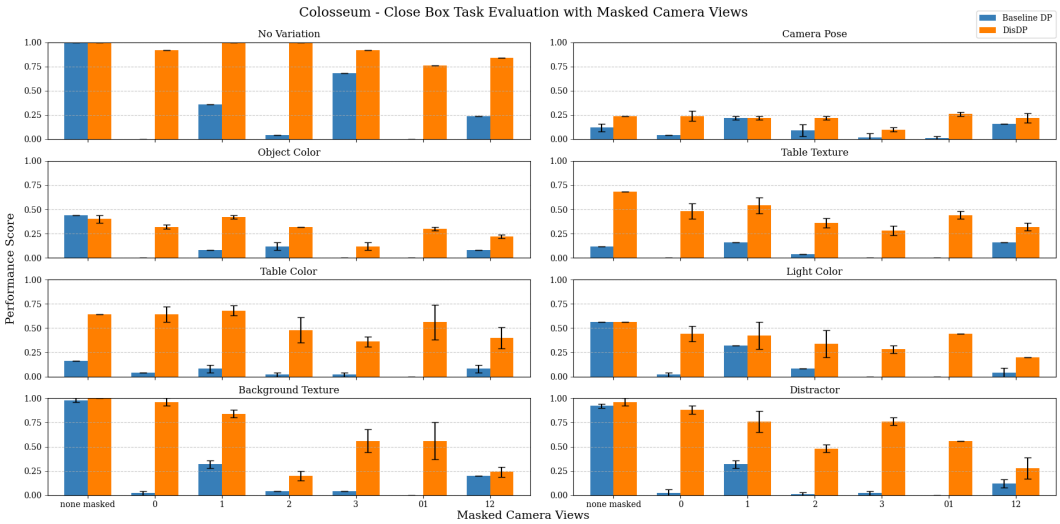

Figure 7: Close box Variations

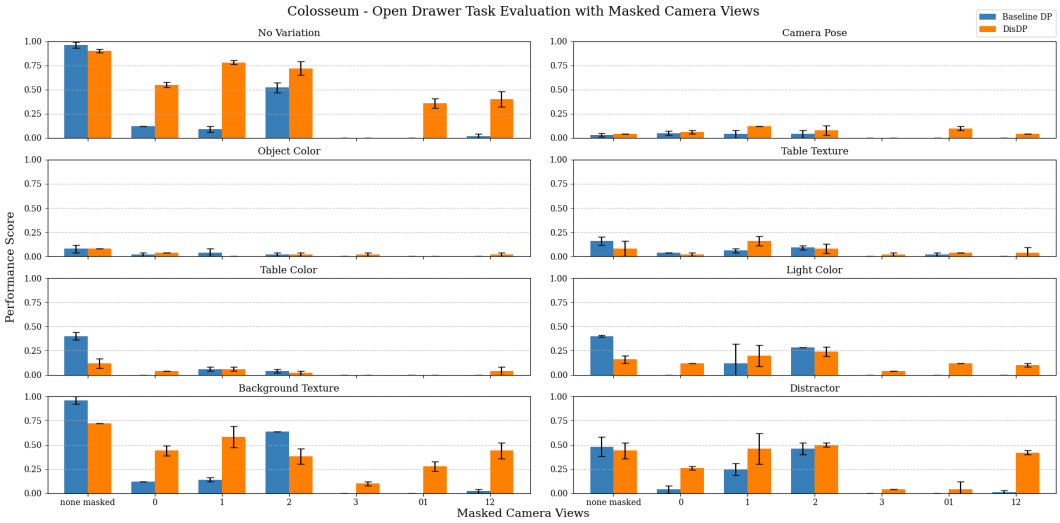

Figure 8: Open Drawer Variations

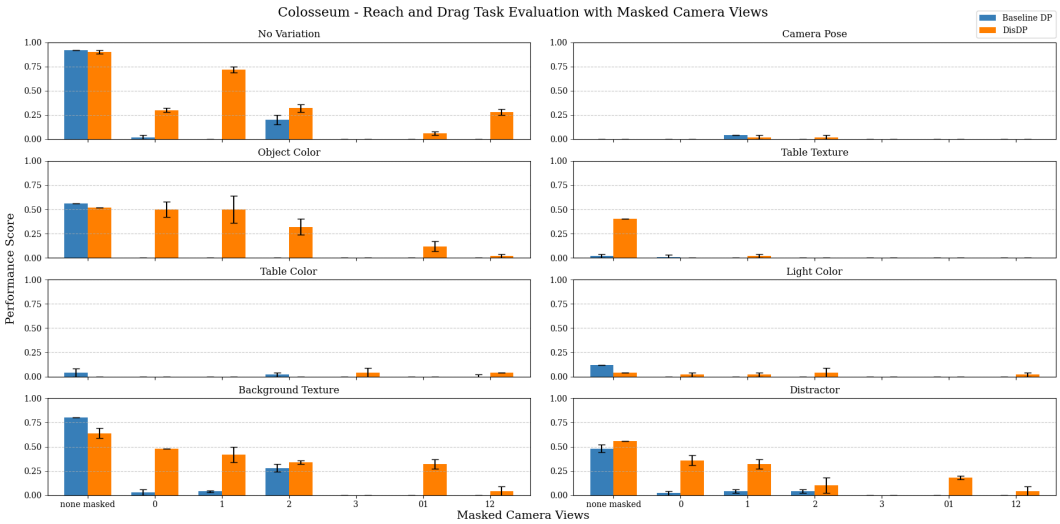

Figure 9: Reach and Drag Variations

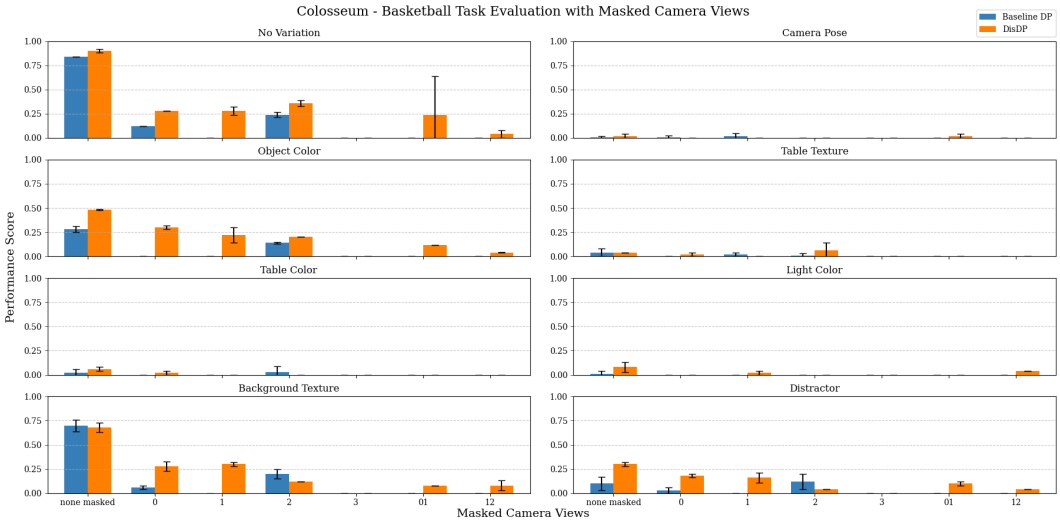

Figure 10: Basketball Task Variations